Voice: The New UI for Mobile Devices

Jan Sedivy

EUROPEAN COMPUTER SCIENCE SUMMIT – ECSS 2010

Fred Jelinek

During 21 years at IBM Research and nearly two decades at Johns Hopkins, he has pioneered the statistical methods that enable modern computers to understand spoken language.

"He envisioned applying the mathematics of probability to the problem of processing speech and language," said Sanjeev Khudanpur, a Johns Hopkins associate

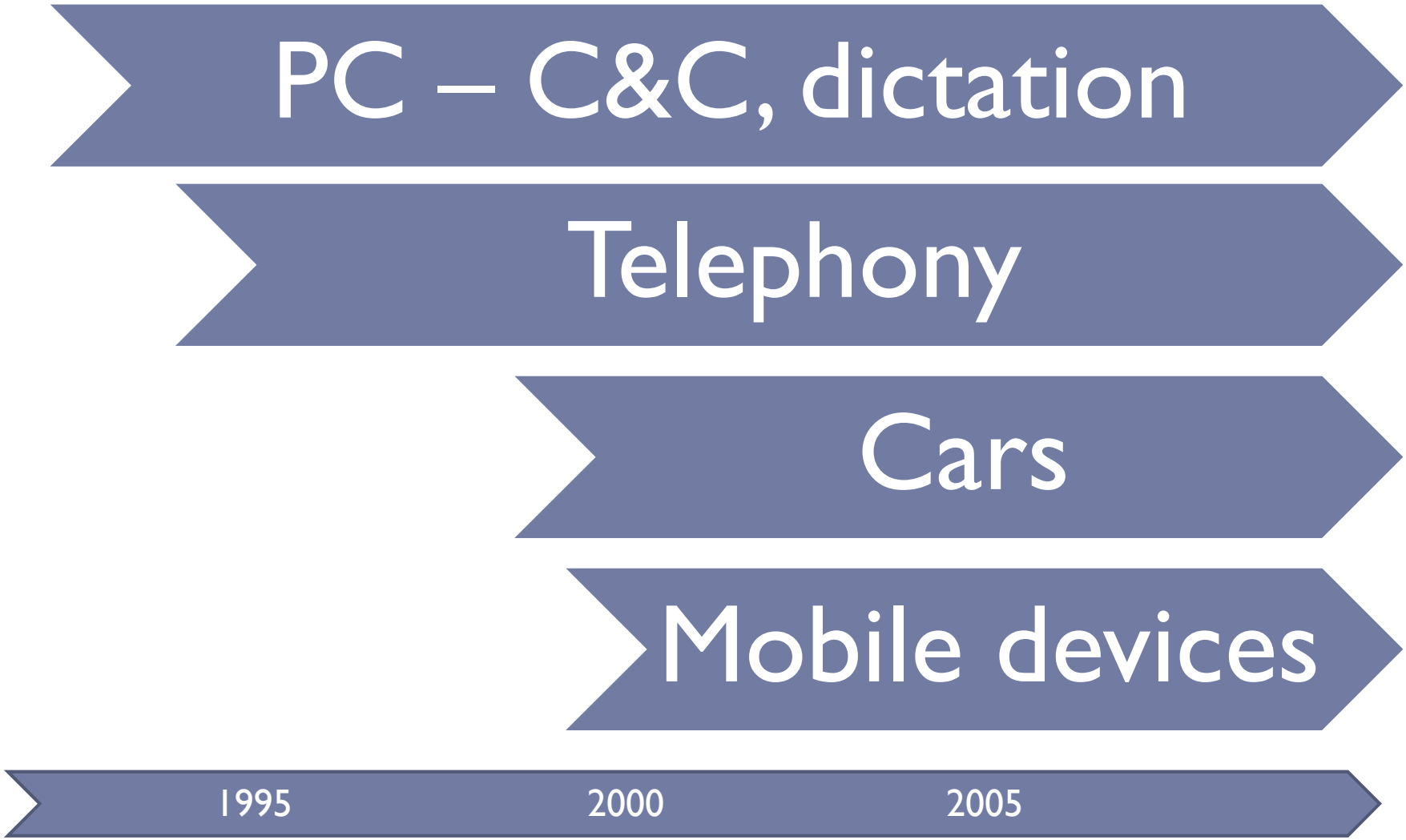
WHY SPEECH RECOGNITION?

Speech recognition areas

Command control, digit dictation	Creation of texts, dictation	Interactive voice response
Automotive speech recognition	Mobile telephony	Voice search

Speech is the most natural way we communicate

The main areas in time perspective

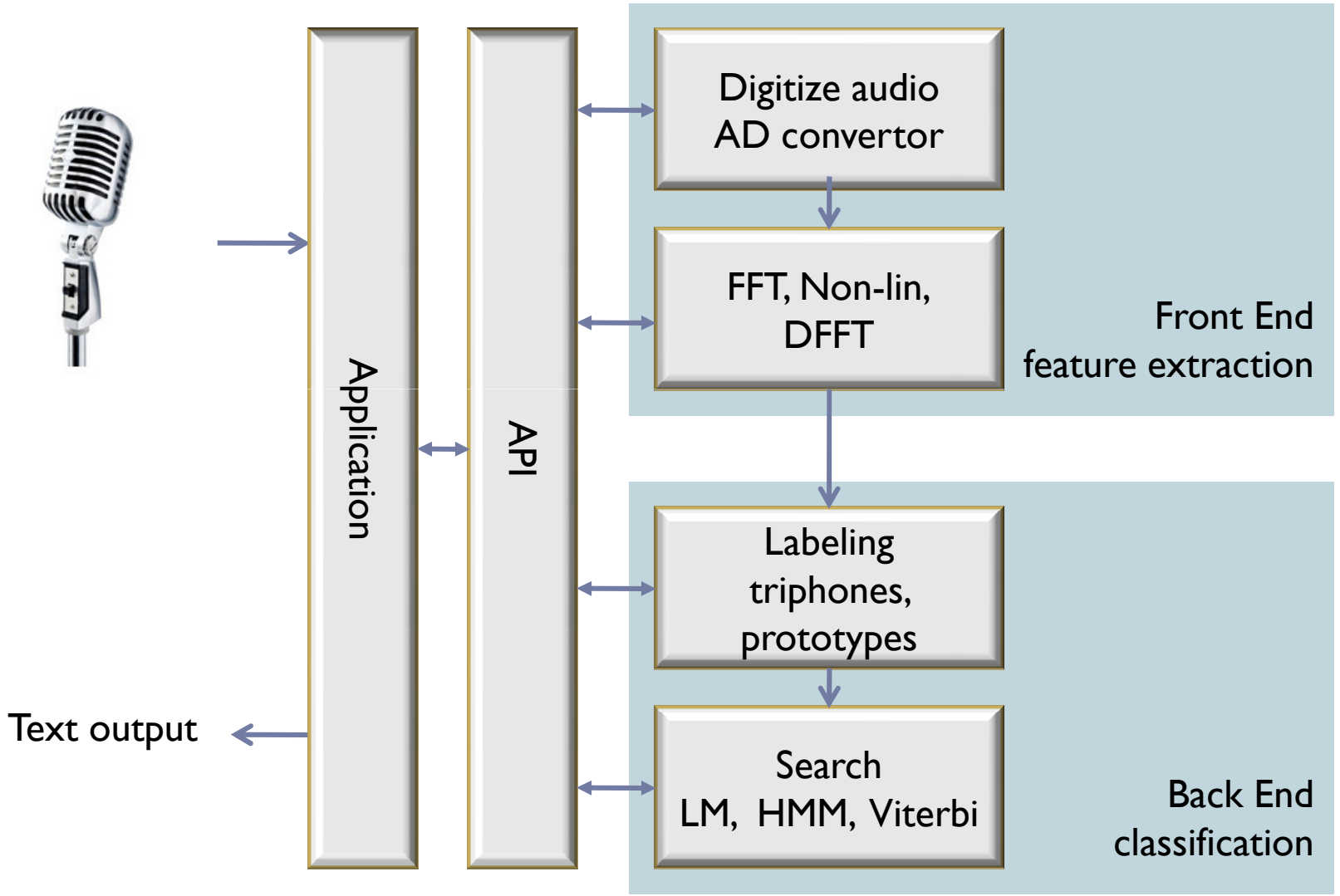


Little more history

- 1993--IBM launches the IBM Personal Dictation System (IPDS)-OS/2, IBM PC, custom audio adapter card
- 1996 VoiceType (Win 95, Netscape, dictation of office document, isolated words, email, …)
- 1996 - Nuance deployed its first commercial speech application
- 1997 Dragon Systems unveiled its Naturally Speaking
- 1999 VoiceXML
- 2000 Telephony applications
- 2002 enabling car control (control car equipment, make a phone call, select music, dictate address to navigation)
- 2003 Microsoft includes speech to Office 2003
- 2007 by the growth of mobile phones/devices
- 2008 Google launches speech to Search iPhone
- 2009 - Nuance Acquires IBM's patents Speech Technology rights

HOW SPEECH RECOGNITION WORKS

Speech recognition – high level



APPLICATIONS DEVELOPMENT CHRONOLOGICALLY

IBM speech recognition – the early days

- Large vocabulary, dictation (1990…)
- Office correspondence task – Tangora
- Written in Fortran
- IBM RISC System/6000, AIX, Tangora

Albert Tangora (July 2, 1903 – April 7, 1978) set the world speed record for sustained typing on a manual keyboard for one hour, 147 words per minute, on October 22, 1923.

How to get reco running on PC -1994

Front End	• Add-on board with ASIC • Integer version on CPU
Hierarchical labeler	• Input - 39 dim cepstrum coeffs feature vector each 10 ms • Output - 100 most likely prototypes out of 30k, diagonal Gaussians
Search	• Statistical LM – high compression, log, • Viterbi search, Hidden Markov Models

How get reco running on Embedded 1999

Easy Port to Embedded

- Resource efficient speech recognition engine
- Written in C/C++
- Integer implementation, GCC compiler
- Simple API to customize for any platform

Basic reco

- Grammar support for command control applications
- Special emphasis on digit recognition
- Robust front end for noisy environments

Cars applications:

- Command control
- Digit and name dialing
- Navigation control
- On-board entertainment control

MOBILE DEVICES

Smartphone sales grow

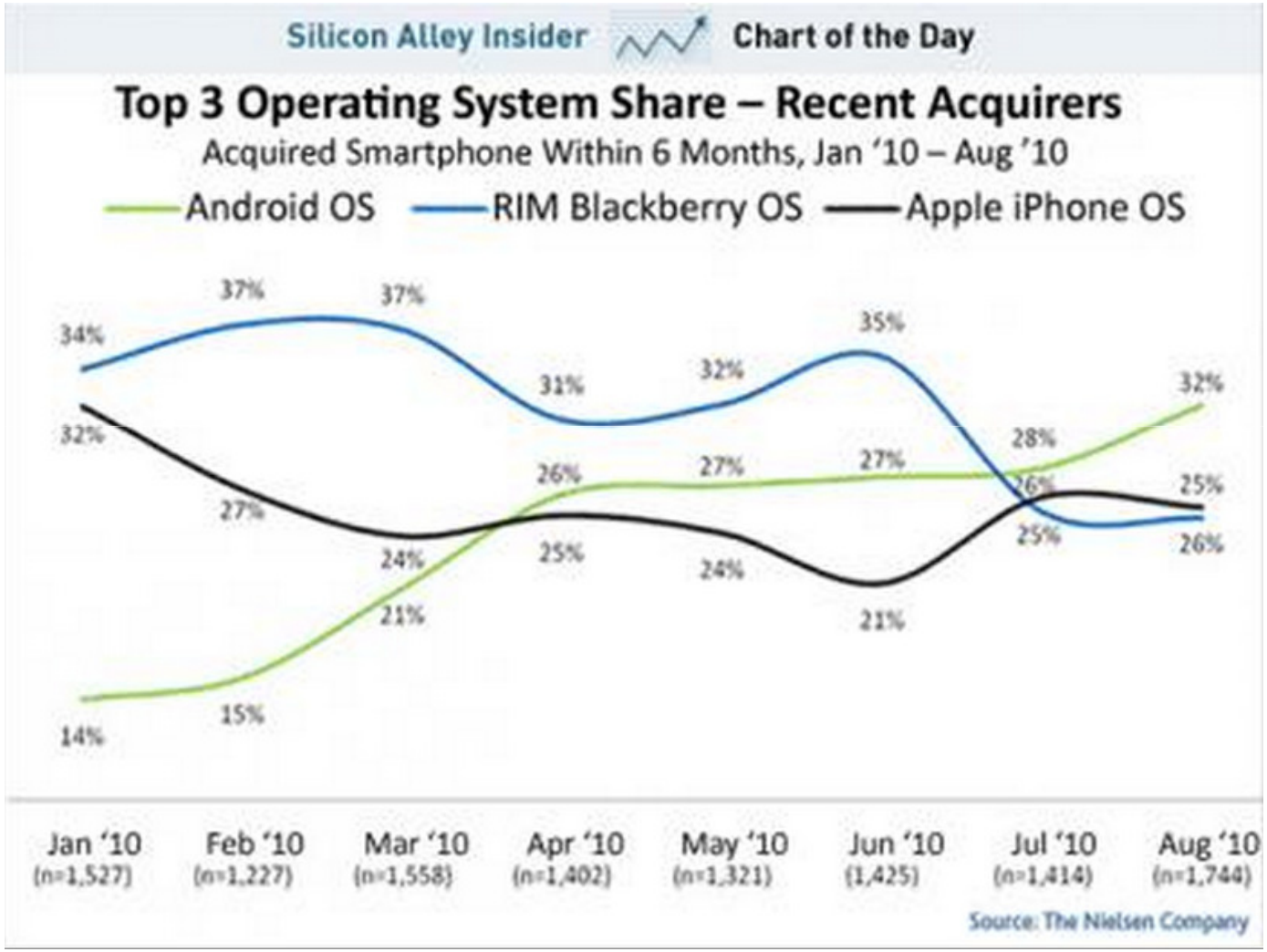
- 2008 1.211billion cell phones,
- 2009 1.222 billion a 0.9 down percent from 2008

- 2009 4Q surge in Smartphone sales to 340 million an 8.3 percent gain over 4Q of 2008,

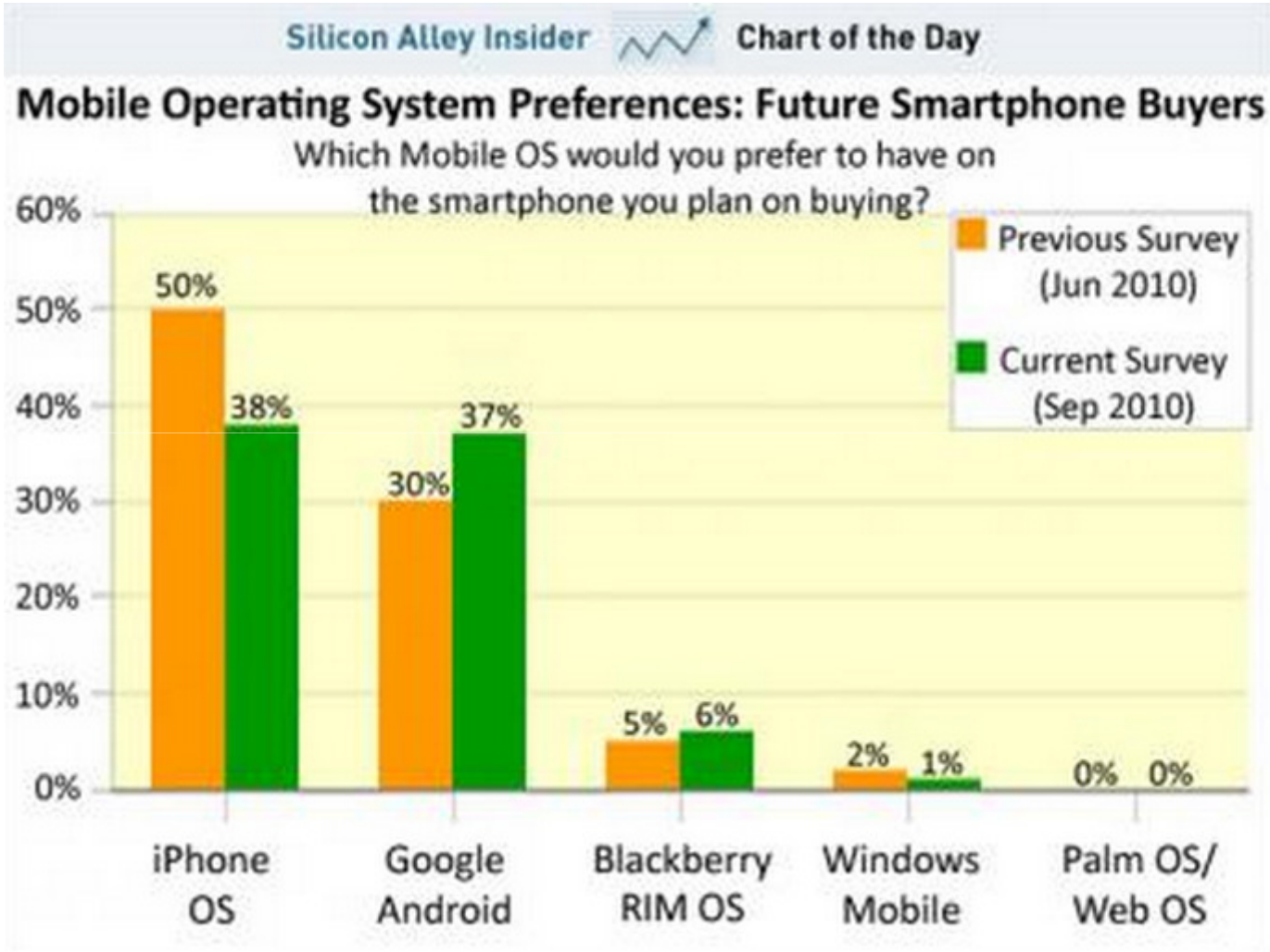
- Mobile phone sales 314.7 mill. ,1Q 2010,17% to 2009,
- Smartphone 54.3 mil., in 1Q 2010, 48.7% up of 2009.

according to Gartner, Inc.

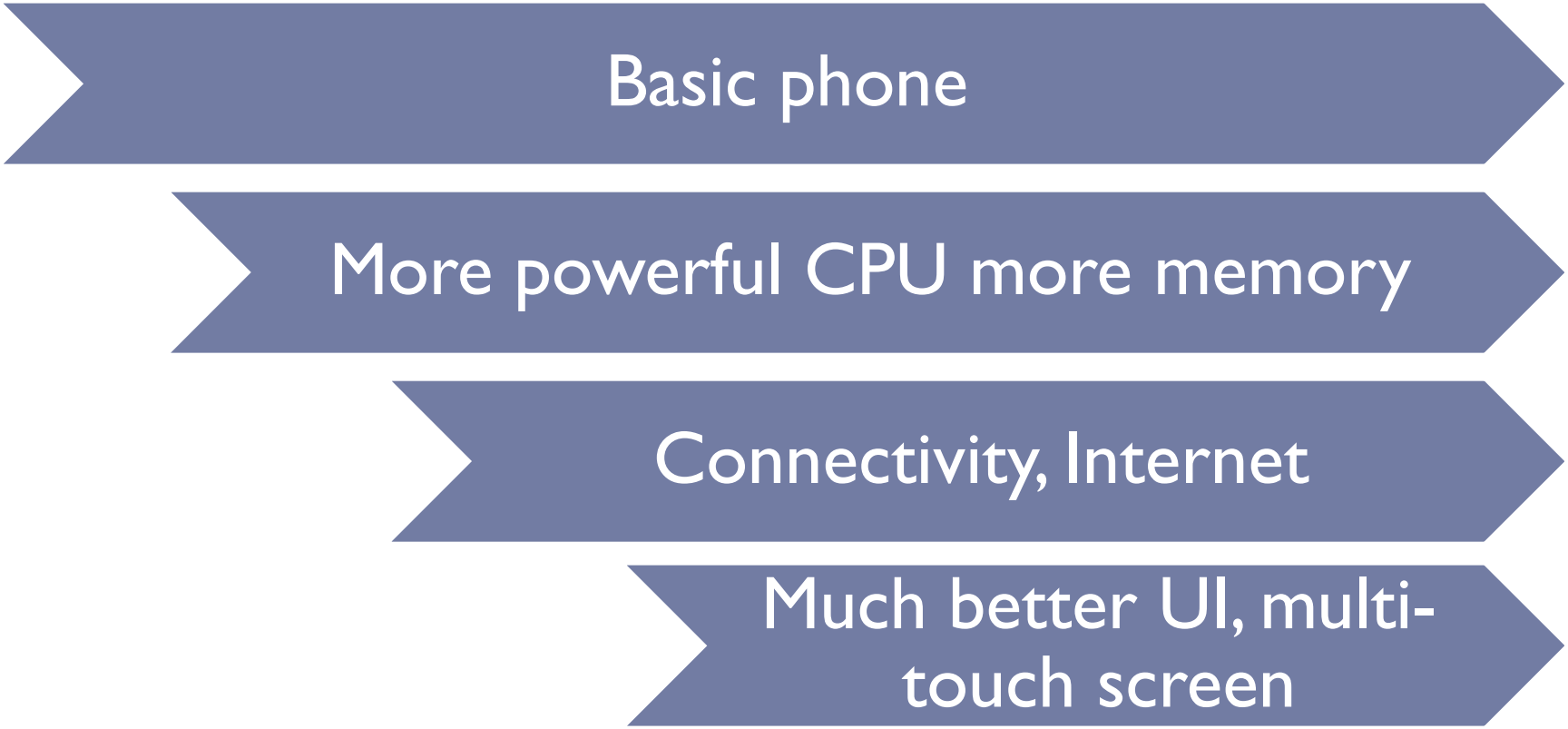
Top 3 operating systems



Mobile operating system preferences



Factors accelerating better mobile apps



Rapid growth of mobile phones/devices is driving the adoption of speech recognition

Why is reco so important for mobile?

- Small form-factor
- Limited keyboard
- Difficult text entry
- Difficult to navigate
- Small screen
- Small CPU
- Slow, not reliable connectivity (latency)

Speech is fundamentally changing the mobile user experience

Speech reco benefits

Speech is rich

- Speech is much richer then two mouse buttons
- Disambiguation, dialog
- Show me all emails from David about Linux server
- "Call David", David Smith or Stone? Home or cell?

Text entry

- Speech expresses not only text entry but C&C, search, URI entry
- Speech entry is part of the keyboard
- "command box", general source of information

WYSIWYG == What You **Say** Is What You Get

Elements of success

Best accuracy:

- Access to huge content: Internet, YouTube, maps, music, pictures, SMS, email…
- Train on all available data: contact, location names addresses, email, documents content, history, personalization and other sensors: GPS, accelerometers, camera, compass
- Computationally expensive - huge clusters of computers to speed up training

Great UI design:

- speech reco must not introduce any friction to the interface
- keyboard, touch screen, multi-touch, keyboard, speaker, microphone
- OS control, part of the OS, noise reduction, AD converter
- Use all sensors available on the phone to inject extra information to app

Demonstration

- Search Conversion
- SMS dictation
- Czech version

Phone: Nexus One, two mics
OS: Android - Froyo 2.2.1

Future challenges

- Better speech recognition accuracy (noisy conditions, dictation)
- Semantic extraction
- Understanding multiple languages (how would a German native search for an address in France?)
- Automatic language adaptability, accents, meanings, topics
- Learning from multiple sources, (voice, text, video, sensors)
- Better user interfaces, multimodal UIs
- Better integration of speech reco to the new applications

Questions and thank you